Revista Ícone
Revista de Divulgação Científica em Língua Portuguesa, Linguística e Literatura
Volume 10 – Agosto de 2012 – ISSN 1982-7717

DIFICULDADES NA COMPILAÇÃO DE UM *CORPUS* DE LÍNGUA ESPANHOLA

Laura Campos de Borba – UFRGS/PIBIC/CNPq [1]
lauracborba@hotmail.com

RESUMO: Algumas das ferramentas utilizadas pelos pesquisadores da Linguística de *Corpus*, visando a criação de corpora através da coleta de textos da internet, são os navegadores off-line. O objetivo do presente trabalho é apresentar os problemas encontrados durante a coleta de textos da internet para a compilação de um *corpus* de língua espanhola. A metodologia utilizada é o manejo do navegador off-line *HTTrack* para reunir e armazenar textos jornalísticos em espanhol. Nossos primeiros resultados demonstram que a utilização desse programa não fornece garantias de que todo e qualquer texto disponível na internet possa ser obtido.

Palavras-chave: Compilação de *corpus*. Navegadores off-line. Língua espanhola.

Introdução

O presente trabalho é uma das etapas de um projeto[2] maior que possui, entre outros objetivos, verificar, na norma padrão do espanhol, quais elementos lexicais são comuns aos falantes dos países que possuem o espanhol como língua oficial ou como segunda língua, e quais elementos lexicais são utilizados em um país ou região específica. A partir dos resultados da pesquisa, busca-se orientar a respeito da marcação diatópica[3] em um dicionário voltado a estudantes brasileiros de espanhol como L2, no âmbito universitário.

No momento, encontramo-nos na etapa de coleta de textos para a compilação de um *corpus* de língua espanhola. O gênero escolhido é a notícia (MARCUSCHI, 2008), por refletir a norma culta e pelo fato de que o jornal, seu veículo de divulgação, é de grande circulação (cf. PERUZZO 2007, p. 57; ZANATTA 2010, p. 110).

---

[1] Graduanda em Letras pela Universidade Federal do Rio Grande do Sul (UFRGS) e bolsista de Iniciação Científica PIBIC/CNPq. E-mail: lauracborba@hotmail.com

[2] O projeto *Bases teórico-metodológicas para um dicionário monolíngue de espanhol como L2 para estudantes universitários brasileiros* é desenvolvido no Instituto de Letras da Universidade Federal do Rio Grande do Sul (UFRGS), cujo início se deu no ano de 2007.

[3] Em um dicionário, a marcação diatópica é a indicação a respeito da localidade (países ou regiões de um determinado país) na qual um verbete ou uma acepção são utilizados.

Foram coletados jornais de cada um dos vinte[4] países que têm o espanhol como língua oficial e para os dois[5] países que o adotam como segunda língua. Limitamos para 01 (um) o número de periódicos de cada país, pois alguns países, devido à sua pequena extensão territorial, têm um único jornal de circulação nacional.

Todos os jornais correspondem a versões disponibilizadas na internet. A ferramenta escolhida para a sua obtenção é o navegador off-line *HTTrack,* indicado por Sardinha (2004). Esse programa tem a função de realizar o download de páginas da internet (com seus arquivos de áudio, vídeo, imagens e outros) para um computador. No entanto, conforme será desenvolvido posteriormente, esse programa não funcionou totalmente, já que não foi possível realizar o download de todos os textos.

O objetivo do presente trabalho é apresentar os problemas que estão relacionados tanto ao manejo do programa *HTTrack* como à dificuldade em encontrar jornais que, em suas configurações, permitam que o download dos mesmos possa ser realizado.

1 Funcionamento do *HTTrack*

O funcionamento do *HTTrack* comporta três passos. O primeiro é atribuir um nome ao projeto, no espaço *Nome do Projeto*, e escolher, em *Caminho Base*, o local onde serão armazenados os arquivos a serem baixados. Em seguida, clica-se em *Avançar*.

Na tela seguinte, o segundo passo consiste em selecionar a opção *Copiar site(s) da web* e, no campo *Endereço Web (URL)*, digitar o endereço do site do jornal escolhido.

O terceiro passo é selecionar os tipos de arquivos que o programa deve baixar. Para os nossos objetivos, os arquivos que são relevantes são os que contém a extensão de texto *.html*. Ainda na tela apresentada no segundo passo, deve-se clicar no botão *Definir as opções*. Na aba *Filtros*, deve-se trocar todos os sinais "+" pelo sinal "-". Em seguida, ao clicar em *Excluir links*, seleciona-se a opção *todos os links*, no campo *Escolha uma regra*, e clica-se em *Adicionar*. No botão *Incluir links*, no campo *Escolha*

[4] Espanha, Colômbia, Equador, México, El Salvador, Venezuela, Chile, Peru, Guatemala, Costa Rica, Panamá, Cuba, Paraguai, Bolívia, República Dominicana, Nicarágua, Argentina, Uruguai, Honduras e Porto Rico (cf. LIPSKI, 1996).
[5] Estados Unidos e Filipinas (cf. PALACIOS, 2008).

uma regra, seleciona-se a opção *Nomes de pasta contendo*[6]. Em *Escolha uma senha*, deve-se pôr a data do dia que se quer baixar, no formato ANOMÊSDIA, ANO/MÊS/DIA ou ANO-MÊS-DIA, e clicar em *Adicionar*. Em seguida, clica-se em *Ok*, *Avançar* e *Concluir*.

2 Dificuldades na coleta de dados

Classificamos as dificuldades enfrentadas em duas categorias: a primeira está relacionada aos jornais escolhidos e a segunda está relacionada ao programa.

2.1 Problemas inerentes aos jornais

2.1.1 Formatação

Há jornais que, quando armazenados no servidor, utilizam a formatação ANOMÊSDIA (ou uma de suas variações). Um exemplo disso é o jornal *La Tribuna*, de Honduras, que armazena seus arquivos com a formatação *2011/10/25*.

Por outro lado, há jornais cuja formatação não permite que se possa visualizar o modo de armazenamento utilizado ou que se possa verificar uma forma padrão de armazenamento em pastas. Isso prejudica o nosso trabalho, pois é necessário especificar uma pasta para a qual o programa será direcionado e fará o download dos arquivos.

Um exemplo de jornal que não permite uma melhor visualização de sua formatação é o venezuelano *Correo del Ávila*. Nesse jornal, as notícias são apenas numeradas e armazenadas em uma pasta única, não havendo uma distinção das notícias por dia de publicação. Uma notícia da seção de política do dia 31 de outubro, por exemplo, apresenta-se com a formatação http://www.correodelavila.com/*noticia.php?id=2671* (grifo nosso), sem qualquer indicação de critérios que apontem para uma organização de pastas por datas. Diante de casos como esse, não é possível estabelecer uma formatação-chave no campo *Escolha uma senha* do *HTTrack* que direcione o programa.

Um exemplo de jornal que não utiliza uma forma padrão de armazenamento em pastas é o *La Hora*, da Guatemala. Este jornal armazena os arquivos em pastas com os

[6] Os sites da internet são normalmente organizados por pastas no servidor, nas quais estão armazenados seus arquivos. Em sites de jornais, essas pastas geralmente são nomeadas com o ano, o mês e o dia das publicações. Por exemplo, o site *http://www.latribuna.hn/2011/10/30/* armazena seus arquivos na pasta *www.latribuna.hn*, que, por sua vez, contém as edições desse jornal organizadas na pasta *2011*. Nessa pasta, está contida a pasta *10* (referente a outubro) e esta, por sua vez, contém a pasta *30* (referente ao dia 30 desse mês), na qual estão armazenados todos os arquivos de notícias do dia 30 de outubro de 2011.

nomes das suas seções (opinión, etc), como em www.lahora.gt/index.php/*economia*/ economia/empresas/146380-la-navidad-inicia-con-el-gran-desfile-gat-continental. Seria necessário realizar vários downloads, direcionando o *HTTrack* a cada uma das pastas correspondentes às seções, o que tornaria nosso trabalho mais lento. Outro problema que o armazenamento em seções traz é a não distinção das notícias por data. Desse modo, não teríamos o exemplar virtual do dia 25 de outubro do jornal *La Hora*, mas sim várias seções separadas, cada uma com notícias desordenadas de vários dias diferentes.

2.1.2 Troca de site

Um segundo problema que enfrentamos está relacionado especificamente à troca de site do jornal *Diario Expreso*, do Equador. Este jornal possuía o endereço *www.diario-expreso.com* e continha uma formatação ANOMÊSDIA. Após baixarmos alguns exemplares desse periódico, em determinado momento o programa passou a não mais conseguir realizar o download de arquivos. Logo descobrimos o que ocasionara a interrupção: o jornal havia mudado de site, adquirindo o endereço *http://www.eldiario.com.ec/*. O modo de armazenamento dos seus arquivos também havia mudado, passando a ser por seções.

2.1.3 Bloqueio da ação do *HTTrack*

Outro problema que enfrentamos foram os jornais que impediram que o *HTTrack* realizasse o download de seus arquivos. Como exemplo, temos a versão uruguaia do jornal *El País*, disponível em *http://www.elpais.com.uy*. Este jornal armazena suas notícias através da formatação */111031/* (ANOMÊSDIA). Porém, o *HTTrack* não baixava nada além do *index*, a página inicial do jornal. Após várias tentativas, elaboramos a hipótese de que esse site estaria acompanhado de algum programa que bloqueasse a ação do *HTTrack*. Através de pesquisas sobre o assunto, descobrimos que é bastante comum o uso de programas que impedem a ação de hackers em sites da internet. No caso do jornal *El País* do Uruguai, é possível que haja um programa que detecte a presença do *HTTrack* e impeça-o de acessar o banco de dados do jornal (ainda que a ação de navegadores off-line não seja de caráter ilegal).

2.1.4 Download dos jornais de acordo com o dia

No campo do *HTTrack Escolha uma senha*, deve-se colocar a formatação do jornal. Em uma formatação por pastas organizadas de acordo com as datas de

publicação dos jornais, teoricamente seria possível escolher qualquer data (por exemplo, 20111031). Contudo, o que pudemos verificar é que a maioria dos jornais não disponibiliza qualquer edição já publicada; alguns permitem apenas o download do jornal do dia corrente e outros permitem apenas o download do jornal do dia corrente e do dia imediatamente anterior. Como exemplos do primeiro caso, temos os jornais *El Diario* (Bolívia), *La Nación* (Costa Rica), *Granma* (Cuba), *Diario Colatino* (El Salvador), *La Jornada* (México) e *El Nacional* (República Dominicana). Como exemplos do segundo caso, temos os jornais *Crónica* (Argentina), *ABC* (Espanha), *La Raza* (EUA), *La Tribuna* (Honduras) e *La Prensa* (Nicarágua).

Os problemas que esse tipo de restrição gera são a lentidão na compilação do *corpus* e a necessidade de agir com uma disciplina mais rígida, baixando os jornais todos os dias, se quisermos aumentar da maneira mais rápida possível o nosso *corpus*.

2.2 Problemas inerentes ao programa *HTTrack*

O *HTTrack* vem acompanhado de um manual de instruções gerais de uso sobre o programa, em inglês, no menu *Ajuda>conteúdo*, que auxiliam no objetivo original do programa, a navegação off-line. Para coletar textos e formar um *corpus*, são necessárias instruções específicas para que se possam baixar somente arquivos de texto. Sardinha (2004) fornece instruções sobre como utilizar o *HTTrack* para os fins da Linguística de *Corpus*; porém, essas instruções não demonstram clareza e completude suficientes para que o programa possa ser utilizado por um leigo. As instruções de uso do programa aqui apresentadas foram elaboradas com base na leitura de Sardinha (2004) e aprimoradas com detalhes fundamentais descobertos por nosso grupo.

Conclusão

Conforme apresentado no tópico anterior, os problemas por nós elencados têm características bastante singulares, que vão desde as dificuldades na visualização da formatação dos sites até as dificuldades no uso do próprio programa. Causas tão diversas, no entanto, resultam em consequências comuns, tais como: a escolha de jornais que nem sempre são os mais representativos de seus países; e a lentidão no processo de obtenção desses jornais.

Outro aspecto importante a ser considerado é o fato de que, durante as nossas tentativas de compilação do *corpus*, verificamos que não tínhamos conhecimento suficiente da área de Computação, o que contribuiu, em parte, para a lentidão em nosso

trabalho. Diante desse ocorrido, podemos concluir que compilar um *corpus* requer conhecimentos não só da área da Letras, mas também da área da Computação. Esses conhecimentos vão além daqueles próprios de um usuário leigo de um computador.

Referências

LIPSKI, J. **El español de América**. Madrid: Cátedra, 1996.

MARCUSCHI, L. A. **Produção textual, análise de gêneros e compreensão**. São Paulo: Cortez, 2008.

PALACIOS, A. **El español en América:** contactos lingüísticos en hispanoamérica. Barcelona: Ariel, 2008.

PERUZZO, M. S. **Como lidar com os neologismos no texto jornalístico?**. 2007. 137 f. Dissertação (Mestrado em Letras) – Instituto de Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2007.

SARDINHA, T. B. **Linguística de *Corpus***. São Paulo: Manole, 2004.

ZANATTA, F. A **normatividade e seu reflexo em dicionários semasiológicos de língua portuguesa**. 2010. 270 f. Dissertação (Mestrado em Letras) – Instituto de Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2010.